

INTEGRATING AND MINING DISTRIBUTED ENVIRONMENTAL ARCHIVES ON GRIDS

Mikhail Zhizhin, Eric Kihn, Rob Redmon, Alexei Poyda, Dmitry Mishin,
Dmitry Medvedev and Vassily Lyutsarev

ICSU environmental databases

- International Council for Science (ICSU) World Data Centers system founded in 1948
- Fast growing
 - ▣ From increasing number of sensors
 - ▣ From computational models
 - ▣ Mostly as time series = large arrays of numeric data or images
- Distributed and heterogeneous
 - ▣ Text and binary files
 - ▣ SQL databases with diverse schemata

Example: Space Physics Interactive Data Resource

- Solar activity, solar wind data, geomagnetic, ionospheric, cosmic ray, radio-telescope ground observations, telemetry
- Time span: 1933 – present

The screenshot displays the SPIDR (Space Physics Interactive Data Resource) website in a Microsoft Internet Explorer browser. The page title is "SPIDR: Data Categories and Sets". The browser address bar shows the URL: <http://clust1.wdcb.ru/spidr/dataset.do>. The website header includes a navigation menu with "Home", "User Profile", "Data Sets", "Time Interval", and "Data".

The main content area is titled "Data Categories and Sets" and features a sidebar on the left with user status and configuration options. The user is logged in as "vassily1". The "Time Interval" is set to "Feb 17, 1996 - Feb 19, 1996". The "Data Basket" shows 0 parameters and 0 stations. The "SPIDR Tools" section includes links for "Matlab viewer", "Web Services Guide", and "More Info". The "SPIDR Publications" section includes a link to the "Solar-Geophysical Data magazine". The "SPIDR Interfaces" section includes a link to "Guided by Guru".

The main content area is divided into sections for "Index Data", "Station Data", and "Satellite Data". The "Index Data" section is expanded, showing a list of data sets with columns for "Data Sets", "Info", "Metadata", "FTP", "Coverage", and "Native Time".

Data Sets	Info	Metadata	FTP	Coverage	Native Time
AMIE derived Indices	i	📄		global	1 min
Geomagnetic Indices	i	📄	📄	global	1, 3 hr, 1 day
HPI DMSP Data	i	📄		10 satellites	floating: abo
HPI NOAA Data	i	📄		10 satellites	floating: abo
Solar Data	i	📄	📄	global	1 day

The "Station Data" section is also expanded, showing a list of data sets with columns for "Data Sets", "Info", "Metadata", "FTP", "Coverage", and "Native Time".

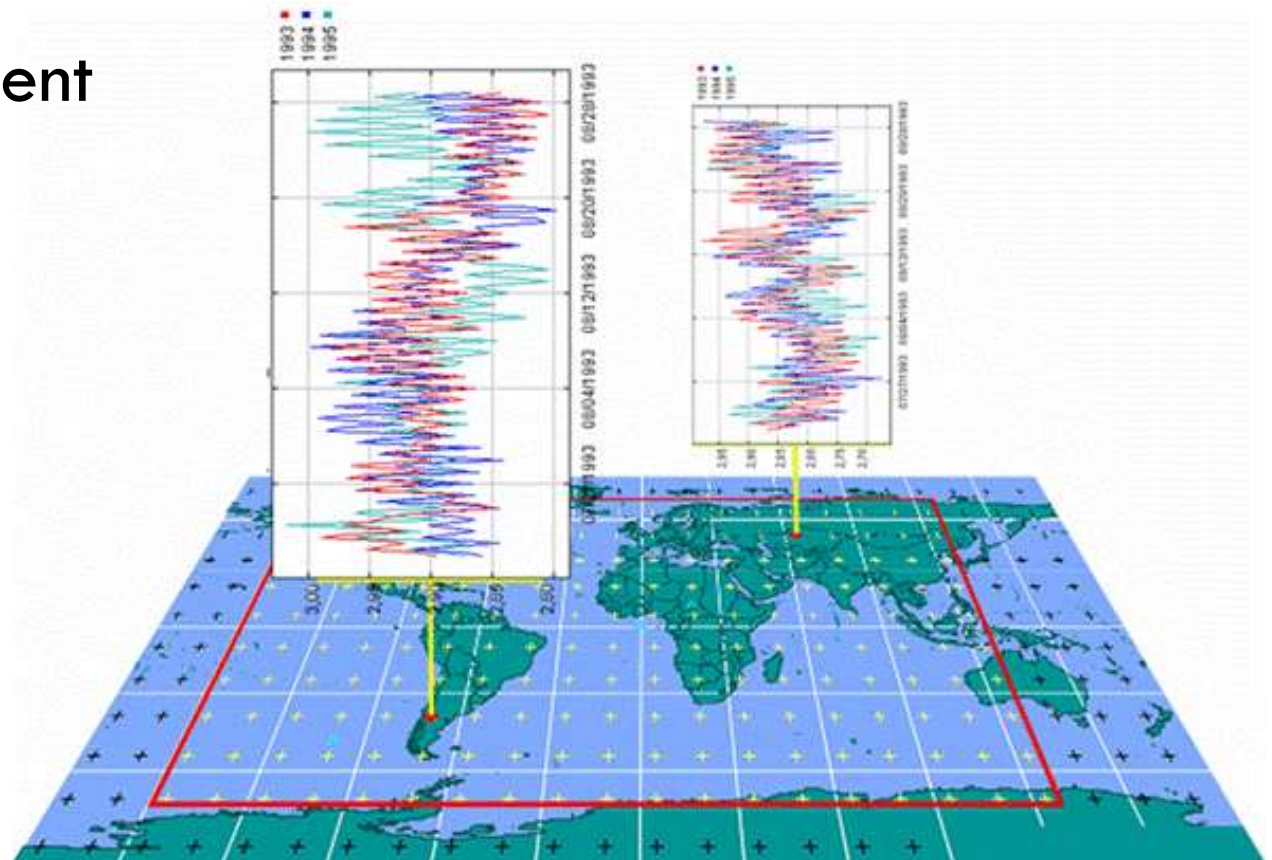
Data Sets	Info	Metadata	FTP	Coverage	Native Time
Cosmic Ray Data (4096 format)	i	📄		120 stations	1 hr
Cosmic Ray Data (general format)	i	📄		39 stations	1 hr
Cosmic Ray Data (preliminary)	i	📄	📄	5 stations	5 min, 1 hr
Geomagnetic Hourly Means	i	📄		229 stations	1 hr
Geomagnetic Minute Means	i	📄		215 stations	1 min
Geomagnetic Yearly Means	i	📄		581 stations	1 month
Ionospheric Data	i	📄		219 stations	floating: 15 n
Radio Solar Telescope Network (RSTN)	i	📄	📄	6 stations	1 sec

The "Satellite Data" section is also expanded, showing a list of data sets with columns for "Data Sets", "Info", "Metadata", "FTP", "Coverage", and "Native Time".

Data Sets	Info	Metadata	FTP	Coverage	Native Time
GOES - Space Environment Monitor	i	📄	📄	N/A satellites	1, 5 min
IMF - Interplanetary Magnetic Field by Minute	i	📄		3 satellites	1, 5 min
IMF OMNI - Interplanetary Magnetic Field by Hour	i	📄		global	1 hr

Example: Climate Reanalysis Project

- $2.5^\circ \times 2.5^\circ \times 6\text{hr} \times 100$ weather parameters
- Time span:
1948 – present



Problem statement

- There is lots of potentially interesting information inside. How to make it useful for ecologists, social scientists, general public?

Environmental scenario search engine

- Common data model
 - ▣ Set of linked multidimensional arrays
 - ▣ Read or append
 - ▣ Selection = hyperslabbing
- Employ existing standards
 - ▣ WSRF for Grid infrastructures
 - ▣ WS-I for more general scenarios
- User-friendly data mining
 - ▣ “Linguistic” terms instead of numeric criteria: **environmental scenarios** like
 - Magnetic storm
 - Atmospheric front

Scenario in fuzzy logic

atmospheric front =
(high V-wind speed) AND
(low pressure) AND
SHIFT (dt=1 day,
(low U-wind speed) AND
(low V-wind speed)) AND
(high pressure)

Temporal Extent

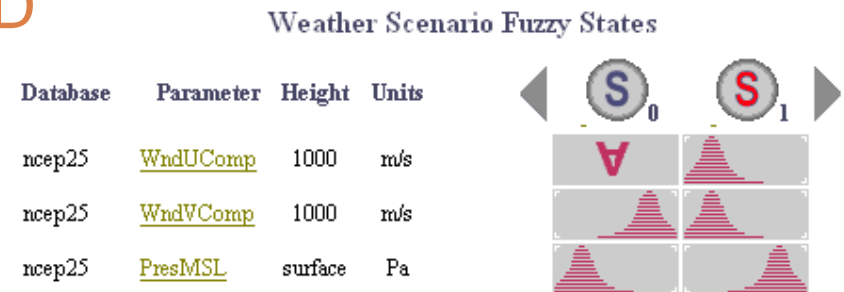
Seasonal time intervals

Date Range: 19490101 to 20051231

Date from, inclusive (year month day): 1995 Jan 1

Date to, inclusive (year month day): 2005 Dec 31

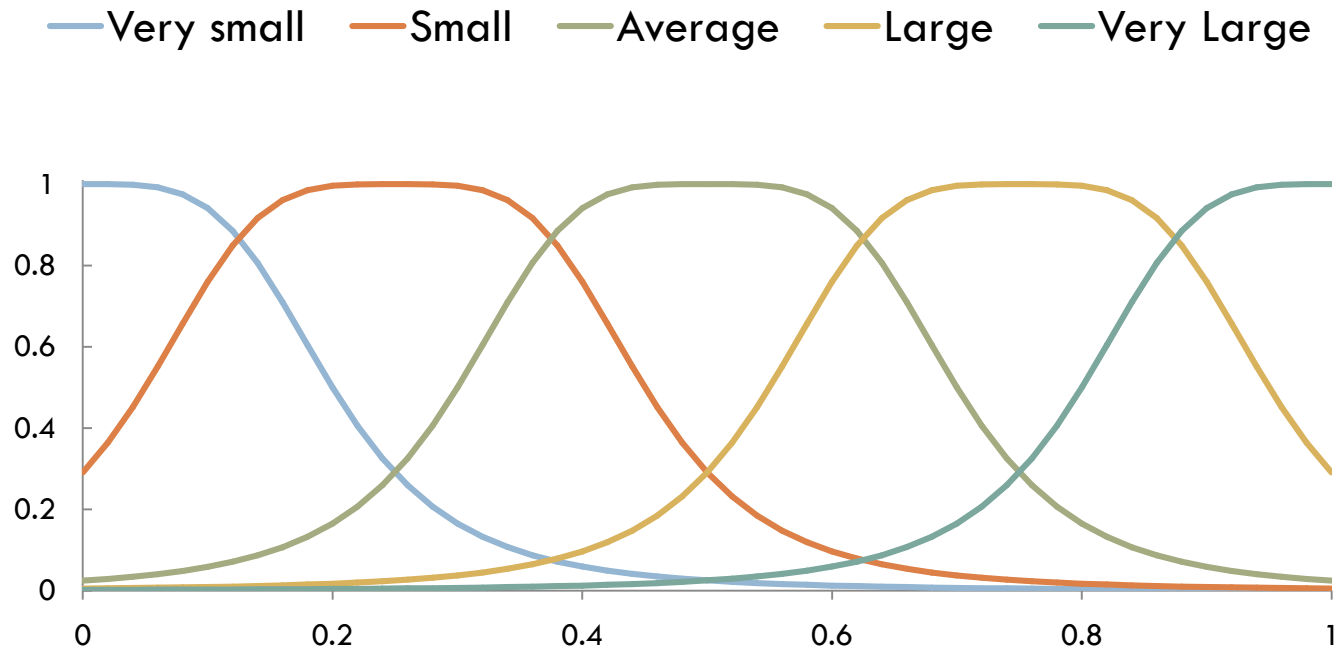
Time window: 1 day



Scenario search Remove current state Clear scenario

Fuzzy conditions

- Fuzzy truth: any value between 0 and 1
- Fuzzy logic: operations on fuzzy truth
- Fuzzy conditions: assign membership functions to linguistic terms

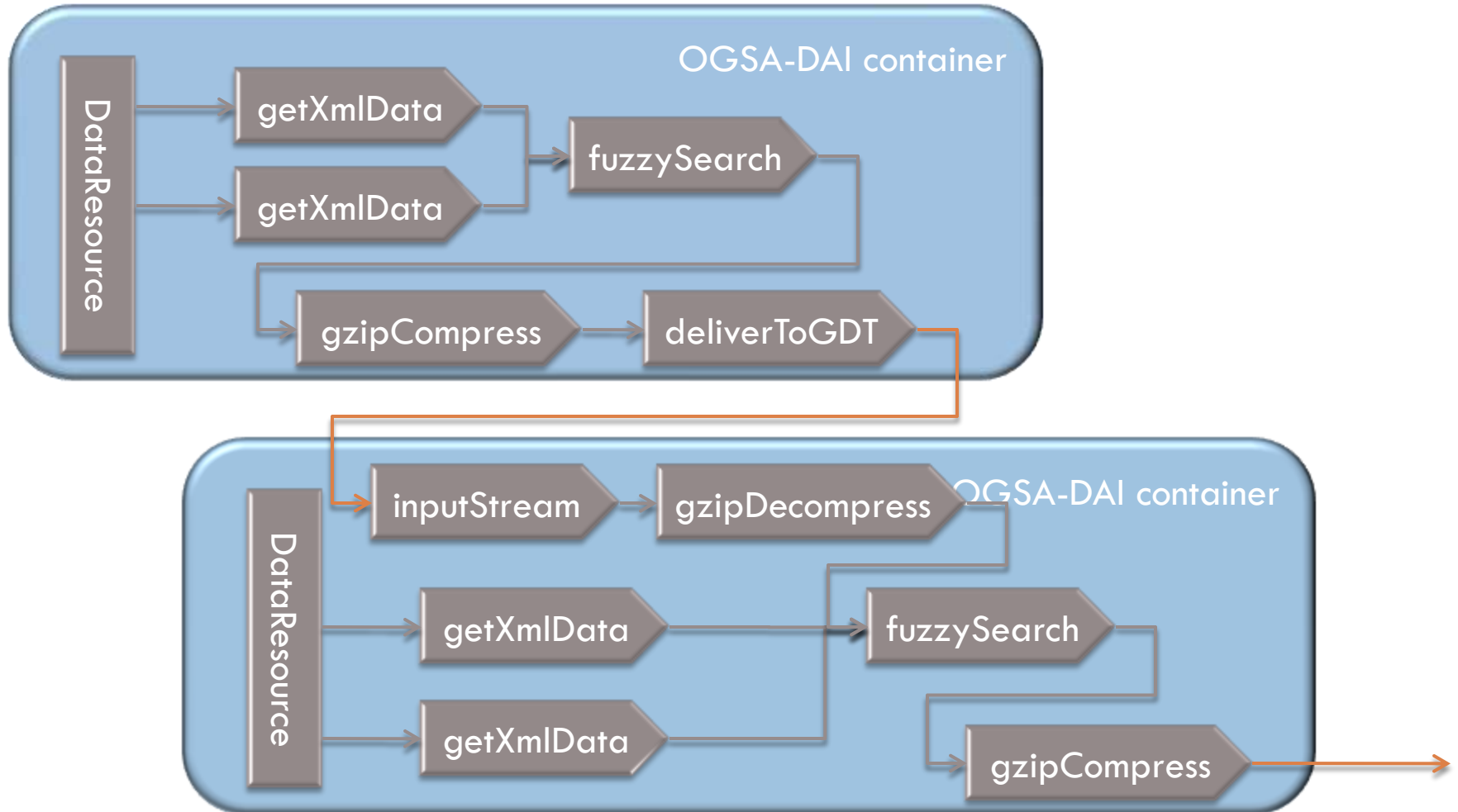


Implementation

- OGSA-DAI: WSRF/ WS-I-compatible web service, extendable, distributed workflow
- EsseDataResource data resource
 - ▣ Sampling and averaging/interpolation of time series
- GetMetaData activity
 - ▣ Enumerates available parameters with their time intervals
- GetXmlData activity
 - ▣ Query data resource and output one time series
- FuzzySearch activity
 - ▣ Perform environmental scenario search logic

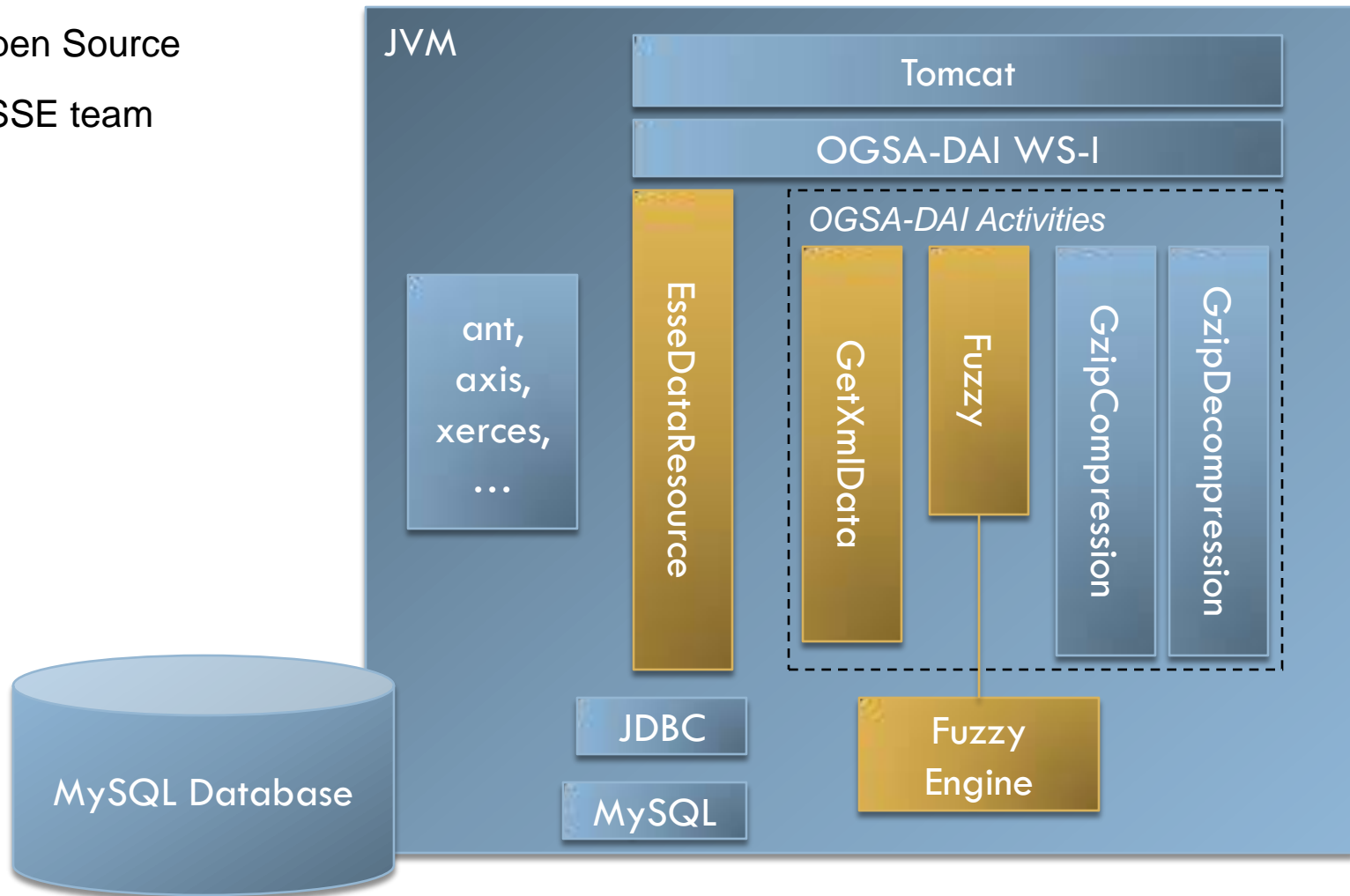
Activity	Representation description	Data size, KB
getNetCdfData	Separate stream with binary NetCDF file	925
getXmlData	Response document with XML. XSD data type: float list	1771
getXmlData+ gzipCompress	Response document with base64 encoded compressed XML	124

Example workflow



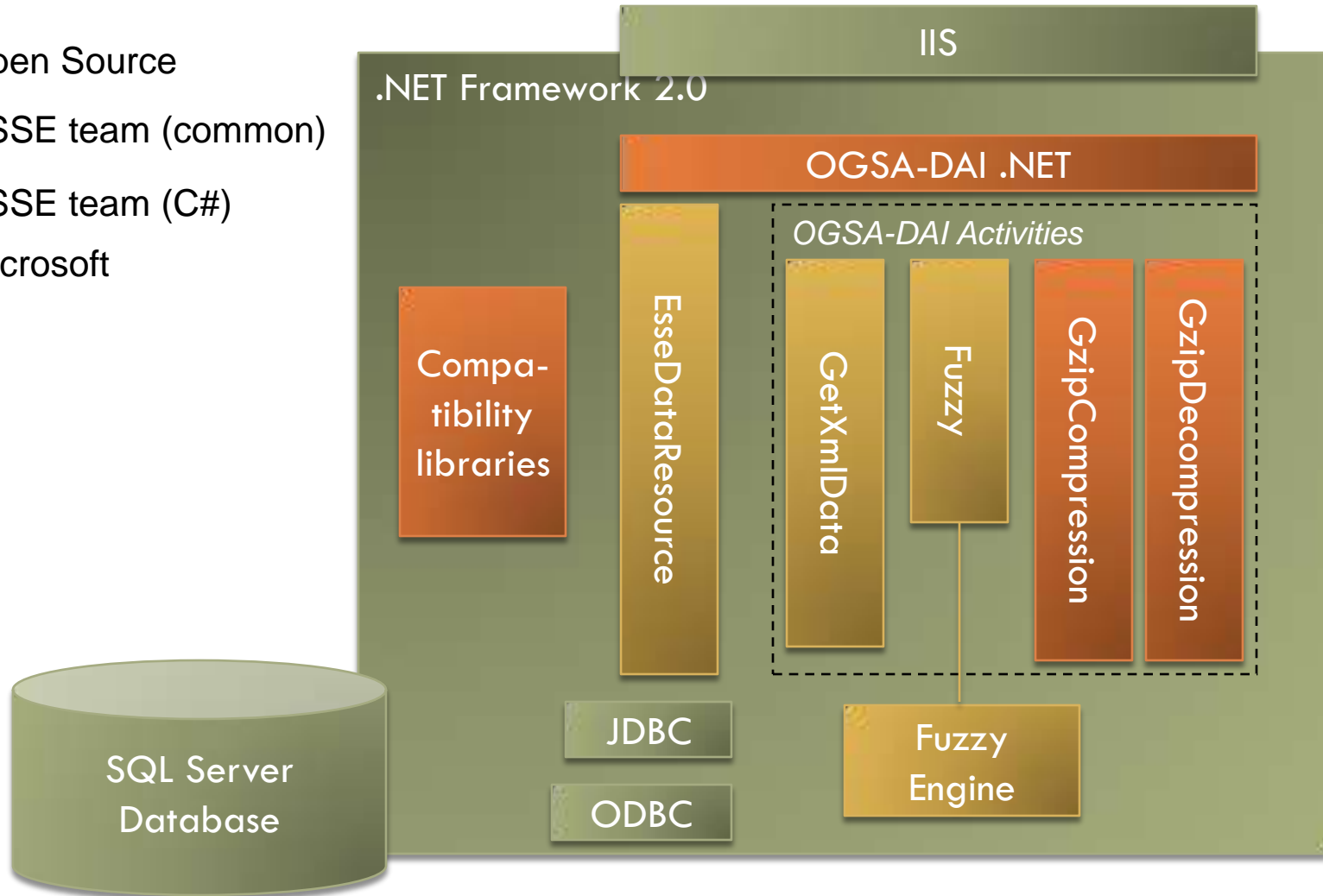
ESSE architecture (Linux platform)

- Open Source
- ESSE team



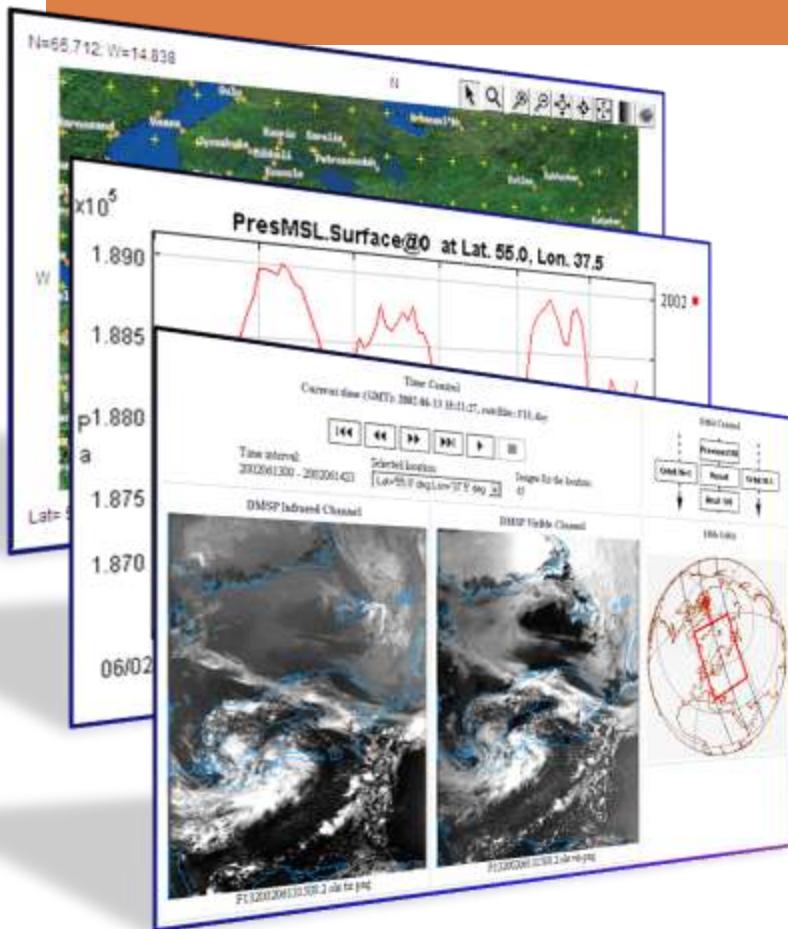
ESSE on a Microsoft platform

- Open Source
- ESSE team (common)
- ESSE team (C#)
- Microsoft



User interfaces

Web portal



NASA World Wind



Conclusions

- OGSA-DAI framework can be used as an interface to archives of environmental data
 - Accessible as web service and on grids
 - More users
 - Interdisciplinary applications
 - Incorporating non-standard data mining algorithms into data access infrastructure